



Contrast Pattern Mining With the T1D Exchange Clinic Registry Reveals Complex Phenotypic Factors and Comorbidity Patterns Associated With Familial Versus Sporadic Type 1 Diabetes

Erin M. Tallon,¹ Maria J. Redondo,²
Chi-Ren Shyu,^{1,3,4} Danlu Liu,³
Katrina Boles,¹ and
Mark A. Clements⁵

Diabetes Care 2022;45:e56–e59 | <https://doi.org/10.2337/dc21-2239>

Scant attention has been paid to evaluating differences in the prevalence of comorbidities and diabetes-related complications in familial versus sporadic type 1 diabetes (1). Knowledge gains in this area could advance the development of risk prediction tools and tailored interventions for preventing or delaying onset of comorbidities or diabetes-related complications in high-risk patient subgroups.

To address this gap, we applied a computationally optimized, exploratory data mining algorithm to the T1D Exchange Clinic Registry (2). For the first time in a large U.S.-based cohort, we assessed demographic and phenotypic factors and comorbid conditions for associations with familial (i.e., having an affected first-degree relative) or sporadic (i.e., having no family history of type 1 diabetes) disease.

The T1D Exchange Clinic Registry is a deidentified, publicly available data set comprising 34,013 adult and pediatric participants who received routine clinical care at 83 U.S.-based endocrinology practices between July 2007 and April 2018 (3). We analyzed participants with a family history of type 1 diabetes involving a first-degree relative, i.e., father ($n = 1,464$), mother ($n = 818$), sibling/twin

($n = 1,882$), and/or child ($n = 228$) (total $n = 3,941$) or no family history of type 1 diabetes ($n = 12,291$). Excluding participants >50 years old resulted in a relatively balanced distribution of age and diabetes duration across both subgroups.

A contrast pattern mining algorithm detects significant differences in the frequencies of attributes across two patient subgroups. We used our validated algorithm to discover individual and co-occurring characteristics that were documented significantly more frequently in familial versus sporadic type 1 diabetes. Here, we refer to these characteristics as “patterns” or “feature patterns.” Our algorithm returns feature patterns consisting of one, two, or three elements. Individual elements are synonymous with individual characteristics.

Metrics used in feature pattern analysis include support, growth, and confidence (4,5). Support is the proportion of individuals in a subgroup who are associated with a given feature pattern. Growth is a support ratio between subgroups. Confidence corresponds to the statistical concept of positive predictive value. We used Fisher exact tests to calculate the statistical significance of each pattern ($P < 0.05$) and the Benjamini-Hochberg

(BH) procedure to control for false discovery (false discovery rate of 0.1).

Of 16,232 individuals who met inclusion criteria, 24.3% ($n = 3,941$) had an affected first-degree relative. Median age of familial cases was 18 (interquartile range [IQR] 15, 27) years; for sporadic cases, median age was 18 (IQR 15, 23) years ($P = 0.05$). Median diabetes duration in familial cases was 10 (IQR 6, 16) years; in sporadic cases, median diabetes duration was 9 (IQR, 6, 14) years ($P < 0.001$). Median age at diagnosis was 8 (IQR 4, 12) years in both subgroups ($P = 0.002$). Mean (\pm SD) hemoglobin A_{1c} (HbA_{1c}) for familial cases was $8.4 \pm 1.3\%$ (68.7 ± 14.7 mmol/mol); for sporadic cases, mean HbA_{1c} was $8.3 \pm 1.2\%$ (66.72 ± 13.2 mmol/mol) ($P < 0.001$).

We discovered 590 feature patterns that met a minimum prevalence threshold of 1% in at least one subgroup. After controlling for false discovery, 265 patterns retained statistical significance. These included 29 single-element patterns, 103 two-element patterns, and 133 three-element patterns (Table 1).

Conditions that were significantly enriched in familial type 1 diabetes included hypertension, hyperlipidemia/

¹Institute for Data Science and Informatics, University of Missouri, Columbia, MO

²Texas Children's Hospital, Baylor College of Medicine, Houston, TX

³Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO

⁴School of Medicine, University of Missouri, Columbia, MO

⁵Division of Endocrinology and Diabetes, Children's Mercy Hospital, Kansas City, MO

Corresponding author: Mark A. Clements, maclements@cmh.edu

Received 27 October 2021 and accepted 27 December 2021

© 2022 by the American Diabetes Association. Readers may use this article as long as the work is properly cited, the use is educational and not for profit, and the work is not altered. More information is available at <https://www.diabetesjournals.org/journals/pages/license>.

Table 1—Enrichment of phenotypic characteristics and comorbid conditions in familial and sporadic type 1 diabetes

Feature pattern	Enriched subgroup*	Support: enriched subgroup (%)	Growth: enriched subgroup	Confidence: enriched subgroup	Nonenriched subgroup	Support: nonenriched subgroup (%)	Growth: nonenriched subgroup	Confidence: nonenriched subgroup	P value
One-element feature patterns									
No documented comorbidities	Sporadic	27.28	1.35	0.81	Familial	20.15	0.19	0.19	1.08E-19
Hypertension	Familial	12.00	1.46	0.32	Sporadic	8.24	0.68	0.68	4.00E-12
Asian	Sporadic	1.64	3.08	0.91	Familial	0.53	0.32	0.09	1.52E-08
Non-Hispanic Black	Familial	6.29	1.55	0.33	Sporadic	4.07	0.67	0.67	2.03E-08
Hyperlipidemia/									
dyslipidemia	Familial	21.52	1.22	0.28	Sporadic	17.57	0.72	0.72	4.49E-08
Atherosclerosis	Familial	1.14	3.26	0.51	Sporadic	0.35	0.31	0.49	5.53E-08
RMV disorder	Familial	9.06	1.36	0.30	Sporadic	6.68	0.70	0.70	1.07E-06
Diagnosis age 0–4 years	Familial	29.21	1.15	0.27	Sporadic	25.30	0.73	0.73	1.53E-06
Erectile/sexual dysfunction	Familial	1.50	2.12	0.40	Sporadic	0.71	0.47	0.60	1.57E-05
Gastroesophageal reflux disease	Familial	3.15	1.60	0.34	Sporadic	1.97	0.66	0.66	3.31E-05
Substance abuse disorder	Familial	1.22	2.11	0.40	Sporadic	0.58	0.47	0.60	9.69E-05
Neuropathy	Familial	4.16	1.45	0.32	Sporadic	2.87	0.68	0.68	1.09E-04
Diagnosis age ≥26 years	Familial	4.47	1.42	0.31	Sporadic	3.15	0.69	0.69	1.38E-04
Nephropathy	Familial	4.52	1.36	0.30	Sporadic	3.31	0.70	0.70	5.74E-04
Insomnia	Familial	1.02	2.05	0.40	Sporadic	0.50	0.49	0.60	6.46E-04
Depression	Familial	11.70	1.18	0.27	Sporadic	9.93	0.73	0.73	1.78E-03
Anemia	Familial	1.62	1.62	0.34	Sporadic	1.00	0.66	0.66	1.97E-03
Diagnosis age 13–18 years	Sporadic	14.41	1.15	0.78	Familial	12.51	0.22	0.22	2.59E-03
ADHD	Familial	7.71	1.20	0.28	Sporadic	6.44	0.72	0.72	6.21E-03
Diagnosis age 5–9 years	Sporadic	34.27	1.07	0.77	Familial	32.00	0.23	0.23	8.95E-03
Thyroid disorder	Familial	21.31	1.10	0.26	Sporadic	19.40	0.74	0.74	9.53E-03
Diagnosis age 10–12 years	Sporadic	18.84	1.11	0.78	Familial	17.03	0.22	0.22	1.07E-02
Allergy	Familial	5.33	1.23	0.28	Sporadic	4.34	0.72	0.72	1.11E-02
Sleep apnea syndrome	Familial	1.22	1.54	0.33	Sporadic	0.79	0.65	0.67	1.50E-02
Constipation	Familial	1.73	1.43	0.31	Sporadic	1.20	0.69	0.69	1.63E-02
Hispanic or Latino	Sporadic	8.92	1.16	0.78	Familial	7.71	0.22	0.22	1.89E-02
Overweight/obesity	Familial	4.95	1.20	0.28	Sporadic	4.13	0.72	0.72	3.09E-02
Asthma	Familial	6.06	1.17	0.27	Sporadic	5.17	0.73	0.73	3.50E-02
Diagnosis age 19–25 years	Familial	4.77	1.18	0.28	Sporadic	4.03	0.72	0.72	4.49E-02

Continued on p. e58

Table 1—Continued

Feature pattern	Enriched subgroup*	Support: enriched subgroup (%)	Growth: enriched subgroup	Confidence: enriched subgroup	Nonenriched subgroup	Support: nonenriched subgroup (%)	Growth: nonenriched subgroup	Confidence: nonenriched subgroup	P value
Selected two- and three-element feature patterns									
Hyperlipidemia/dyslipidemia and hypertension	Familial	6.95	1.59	0.34	Sporadic	4.36	0.66	0.66	4.07E-10
No documented comorbidities and diagnosis age 5–9 years	Sporadic	9.15	1.46	0.82	Familial	6.27	0.18	0.18	6.46E-09
RMV disorder and hyperlipidemia/dyslipidemia	Familial	4.95	1.58	0.34	Sporadic	3.14	0.66	0.66	2.71E-07
RMV disorder and hypertension	Familial	3.88	1.68	0.35	Sporadic	2.31	0.65	0.65	3.17E-07
Hyperlipidemia/dyslipidemia and hypertension and RMV disorder	Familial	2.84	1.81	0.37	Sporadic	1.57	0.63	0.63	1.01E-06
No documented comorbidities and diagnosis age 13–18 years	Sporadic	4.38	1.55	0.83	Familial	2.82	0.17	0.17	8.52E-06
No documented comorbidities and diagnosis age 10–12 years	Sporadic	5.16	1.46	0.82	Familial	3.53	0.18	0.18	2.00E-05
Nephropathy and hypertension	Familial	2.64	1.64	0.34	Sporadic	1.61	0.66	0.66	6.01E-05
Nephropathy and hyperlipidemia/dyslipidemia	Familial	2.54	1.65	0.35	Sporadic	1.54	0.65	0.65	7.31E-05
Diagnosis age 5–9 years and RMV disorder	Familial	3.17	1.53	0.33	Sporadic	2.07	0.67	0.67	1.26E-04
Neuropathy and hyperlipidemia/dyslipidemia	Familial	2.51	1.62	0.34	Sporadic	1.55	0.66	0.66	1.36E-04
Depression and hypertension	Familial	2.51	1.55	0.33	Sporadic	1.62	0.67	0.67	4.80E-04
No documented comorbidities and Hispanic or Latino	Sporadic	2.87	1.49	0.82	Familial	1.93	0.18	0.18	1.11E-03

Continued on p. e59

Table 1—Continued

Feature pattern	Support: enriched subgroup (%)	Growth: enriched subgroup	Confidence: enriched subgroup	Nonenriched subgroup	Support: nonenriched subgroup (%)	Growth: nonenriched subgroup	Confidence: nonenriched subgroup	P value
-----------------	--------------------------------	---------------------------	-------------------------------	----------------------	-----------------------------------	------------------------------	----------------------------------	---------

RMV disorder and thyroid disorder

4.78E-03

Two categories of results were used: 1) one-element feature patterns and 2) two- and three-element feature patterns. *P* values were obtained using Fisher exact tests. False discovery resulting from multiple-hypothesis testing was controlled using the BH procedure (false discovery rate, 0.1). Results in both categories (i.e., one-element feature patterns and two- and three-element feature patterns) are sorted by *P* value. Two- and three-element patterns selected for inclusion in this table met the following criteria: 1) confidence was increased relative to related one-element patterns, 2) pattern growth in the enriched subgroup was ≥ 1.4 , 3) pattern support in the enriched subgroup was ≥ 2.5 , and 4) individual pattern elements previously retained significance (as one-element patterns) following use of the BH procedure. ADHD, attention deficit/hyperactivity disorder. *Enriched subgroup is the subgroup in which the feature pattern was documented more frequently.

dyslipidemia, atherosclerosis, retinopathy/maculopathy/vitreopathy (RMV), erectile and sexual dysfunction, gastroesophageal reflux disease, neuropathy, and nephropathy. A higher proportion of individuals with familial disease (vs. sporadic disease) were non-Hispanic Black (6.3% vs. 4.1%). Sporadic type 1 diabetes was more frequently associated with the absence of other medical conditions, Asian race, Hispanic ethnicity, and diagnosis at ages 5–9, 10–12, and 13–18 years.

Hyperlipidemia/dyslipidemia and hypertension, combined, were present for 7.0% of familial cases but for only 4.4% of sporadic cases. Co-occurring RMV and hyperlipidemia/dyslipidemia were documented for 5.0% of familial cases and for 3.1% of sporadic cases.

In contrast to most earlier studies, this study did not exclude patients diagnosed with type 1 diabetes as adults. Across the two subgroups, the difference in median diabetes duration was small (~ 1 year) and mean HbA_{1c} was similar, suggesting that the observed associations cannot be completely explained by the small difference in diabetes duration and HbA_{1c}. An important limitation is that the Registry does not identify whether more than one participant originated from the same family unit; therefore, individual family units may be represented in this analysis more than once.

This study of more than 16,200 individuals in the T1D Exchange Clinic Registry is the largest study to date to evaluate longitudinal health outcomes in individuals with familial versus sporadic type 1 diabetes. Further research is needed to validate the present results in a large population-based cohort.

Acknowledgments. The authors thank all the participants and clinicians of the T1D Exchange Clinic Registry who, through their participation, continue to advance understanding of type 1 diabetes. We also thank Dr. Noah Greifer, Johns Hopkins Bloomberg School of Public Health, for his assistance and valuable comments regarding patient matching methods.

The source of the data used in this analysis is the T1D Exchange, but the analyses, content, and conclusions presented here are solely the responsibility of the authors and have not been reviewed or approved by the T1D Exchange.

Funding. E.M.T. is supported by a grant from the U.S. National Library of Medicine of the National Institutes of Health (5T32LM012410). The T1D Exchange Clinic Registry was originally created through support from the Leona M. and

Harry B. Helmsley Charitable Trust. M.A.C. is supported by an independent grant from the Leona M. and Harry B. Helmsley Charitable Trust (G-2008-04043). The computation for this work was performed on the high-performance computing infrastructure provided by Research Computing Support Services and in part by the National Science Foundation under grant number CNS-1429294 at the University of Missouri, Columbia, MO (<https://doi.org/10.32469/10355/69802>).

The contents are solely the responsibility of the authors and do not necessarily represent the official views of the National Institutes of Health.

Duality of Interest. M.A.C. is the chief medical officer at Glooko. He receives research support from Dexcom and Abbott Diabetes Care. C.-R.S. is a consultant for Curant Health. No other potential conflicts of interest relevant to this article were reported.

Author Contributions. E.M.T. designed the study, performed data cleaning and analysis, and wrote the manuscript. M.J.R. contributed to the discussion pertaining to methodology and reviewed and edited the manuscript. C.-R.S. contributed to the discussion and design of data analytics, reviewed and edited the manuscript, and provided funding for E.M.T., D.L., and K.B. D.L. developed the algorithm used to conduct the analysis, contributed to the discussion, and reviewed and edited the manuscript. K.B. contributed to the discussion, performed data cleaning and analysis, developed data visualizations, and reviewed and edited the manuscript. M.A.C. contributed to the discussion, assisted with data mapping and results interpretation, and reviewed and edited the manuscript. E.M.T. is the guarantor of this work and, as such, had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Prior Presentation. Parts of this study were presented at the National Library of Medicine Annual Informatics Training Meeting, 22–24 June 2020.

References

- Lebenthal Y, Shalitin S, Yackobovitch-Gavan M, Phillip M, Lazar L. Retrospective comparative analysis of metabolic control and early complications in familial and sporadic type 1 diabetes patients. *J Diabetes Complications* 2012; 26:219–224
- Liu D, Baskett W, Beversdorf D, Shyu C-R. Exploratory data mining for subgroup cohort discoveries and prioritization. *IEEE J Biomed Health Inform* 2020;24:1456–1468
- Beck RW, Tamborlane WV, Bergenstal RM, Miller KM, DuBose SN; T1D Exchange Clinic Network. The T1D Exchange clinic registry. *J Clin Endocrinol Metab* 2012;97:4383–4389
- Larose DT, Larose CD. *Discovering Knowledge in Data: An Introduction to Data Mining*. 2nd edition, John Wiley & Sons, Inc., 2014. Accessed 20 April 2021. Available from <https://www.onlinelibrary.wiley.com/doi/book/10.1002/9781118874059>
- Dong G, Li J. Efficient mining of emerging patterns: discovering trends and differences. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining—KDD '99*, 1999. New York, NY, Association for Computing Machinery, p. 43–52